DOCUMENT RESUME

ED 114 103 IR 002 733

AUTHOR Williams, Martha
TITLE The Impact of Machine-Readable Data Bases on Library and Information Services. National Program for Libraries and Information Services Related Paper No. 26.
INSTITUTION National Commission on Libraries and Information Science, Washington, D. C.
PUB DATE Apr 75
NOTE 35p.; For related documents see ED 100 387-97; IR 002 728-34

EDRS PRICE MF-$0.76 HC-$1.95 Plus Postage
DESCRIPTORS Computers; *Data Bases; *Information Retrieval; Information Science; *Information Services; Information Storage; Library Services; Library Technical Processes; National Programs; *Networks; Telecommunication
IDENTIFIERS *National Commission Libraries Information Science; NCLIS

ABSTRACT

The growth and proliferation of machine-readable data bases have created a need to consider the nature of recent developments, the impact of those developments on library and information services, and their relationship with the National Commission on Library and Information Science (NCLIS) program. Data bases may contain information in a variety of forms, may be produced by government or private business, and may be discipline, mission, or problem oriented, or inter- or multi-disciplinary. The availability of such data bases may cause changes in such library activities as journal acquisition and interlibrary loans; or libraries may provide search services, act as intermediaries in preparing searches, or refer people to appropriate information services. The role of NCLIS should be to support education, training, and research in the use of data bases, help expand service to new constituencies, encourage improvement of retrieval systems, promote the use of telecommunications, and provide a basis for networks and data base sharing in all sectors of the information community. (LS)

**********
* Documents acquired by ERIC include many informal unpublished *
* materials not available from other sources. ERIC makes every effort *
* to obtain the best copy available. Nevertheless, items of marginal *
* reproducibility are often encountered and this affects the quality *
* of the microfiche and hardcopy reproductions ERIC makes available *
* via the ERIC Document Reproduction Service (EDRS). EDRS is not *
* responsible for the quality of the original document. Reproductions *
* supplied by EDRS are the best that can be made from the original. *
**********

ED114103

NATIONAL COMMISSION ON LIBRARIES AND INFORMATION SCIENCE
NATIONAL PROGRAM FOR LIBRARIES
AND INFORMATION SERVICES

RELATED PAPER
TWENTY SIX

U.S. DEPARTMENT OF HEALTH,
EDUCATION & WELFARE
NATIONAL INSTITUTE OF
EDUCATION
THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

# THE IMPACT OF MACHINE-READABLE DATA BASES ON LIBRARY AND INFORMATION SERVICES

MARTHA WILLIAMS

DIRECTOR, INFORMATION RETRIEVAL RESEARCH LABORATORY,
UNIVERSITY OF ILLINOIS AT URBANA-CHAMPAIGN

This paper outlines briefly the developments in the area of machine-readable data bases, and to assess the impact of these developments on library and information services; and their relation to the Commission's national program for library and information services.

APRIL, 1975

The views expressed are those of the author and do not necessarily reflect the position or policy of the NCLIS. Though related to the Commission's National Program, papers in this series are not an integral part of the National Program Document.

IR 002 733

TABLE OF CONTENTS


| | Page |
|---|---|
| Introduction | 1 |
| Data Base Origins and Generation | 6 |
| Technical Aspects--Data Base Formats and File Structures | 7 |
| Data Base Characteristics--Criteria for Use | 12 |
| Subject Content | 12 |
| Data Base Search Services | 14 |
| Intermediaries | 19 |
| Data Bases and the National Program | 20 |
| Education and Training | 20 |
| Research | 21 |
| Constituents Served | 23 |
| Resource Locators | 23 |
| Telecommunications | 25 |
| Resource Sharing and Networking | 26 |
| Data Base Generation and Use in the Private Sector | 27 |
| Data Base Problems and Future Trends | 29 |
| References | |
| Related Works | |

# The Impact of Machine-Readable Data Bases on Library and Information Services

## Introduction

The purpose of this paper is: to outline developments in the area of machine-readable data bases; to assess the impact of these developments on library and information services; and to relate them to the National Program for Library and Information Services that has been proposed by the National Commission on Libraries and Information Science.

First of all, one must know what a machine-readable data base is. It is an organized collection of information in machine-readable form. The collected information may be of several types: bibliographic, or bibliographic-related; natural language text; numerical; or representational. An example of a bibliographic data base is the MARC II data base of the Library of Congress or The Chemical Abstracts Service's (CAS) Condensates tapes. The CASIA (Chemical Abstracts Subject Index Alerts) tapes, which contain subject index terms and postings that consist of Chemical Abstracts citation numbers is a bibliographic-related data base because the citation number refers the user to other tapes or hard-copy sources that contain the full bibliographic record for the citation. A natural language text data base would be the text portion of the New York Times Information Bank which contains not the full text of the articles from newspapers but textual summaries or abstracts of the articles. System 50 for State Statutes of Aspen Systems Corporation, an example of a full text data base, contains over 200 million words of statute law in the form of full text. Examples of numeric data bases are numerous but a familiar one would be the U. S. census tapes containing current census data and produced by the United States Bureau of the Census. A data base that contains not alphameric

data but graphic or pictorial representations such as the CAS Registry Structure data base, which contains chemical structures, is referred to as a representational data base. One then can see that there are many types of data bases, containing many types of information which is represented in many ways. In this paper the term data base will refer to bibliographic data bases and no other types unless specifically indicated.

The past decade has seen a considerable number of technological advances and developments in the machine-readable data base field and these have had a decided impact on the types of search services provided to users. Far more data bases exist today than at any time in the past, and far more users are receiving search services from machine-readable data bases than at any time in the past. Aside from advances in the areas of computer technology, storage, and communications, the very simple fact that large numbers of machine-readable data bases, (corresponding to many of the most heavily searched abstracting and indexing services) now exist and can be searched is significant. This was not the case ten years ago. We now have machine-readable data bases in almost all of the major fields of science and technology, as well as data bases covering news articles, legal cases and statutes, drug and poison information, etc. Efforts are underway and certainly more work is needed to generate data bases that would provide community service type information such as consumer, day-care, legal aid, recreational and leisure time activities information, etc. There are hundreds of valuable publicly available data bases and many more private data bases. The problem is making them known, understood and used by researchers and the public at large.

Who produces or generates the data bases? They are produced both by governmental sources and within the private sector. Included in the private sector are profit-making and not-for-profit organizations such as professional societies. Although the government is responsible for the generation of numerous data bases in many cases the actual production work is carried out under contract by not-for-profit or commercial organizations.

Many of the largest and most heavily used data bases were produced by the federal government. Some examples of these are: The MEDLARS (Medical Literature Analysis and Retrieval System) tapes produced by the National Library of Medicine; the MARC II (Machine-Readable Cataloging) tapes produced by the Library of Congress; the ERIC (Educational Resources Information Center) tapes of the National Institute of Education; the DDC Tapes (Defense Documentation Center tapes) of the Department of Defense's Defense Documentation Center; GRA (Government Research Announcements) of the National Technical Information Service (NTIS) and STAR (Scientific and Technical Aerospace Reports) tapes produced by the National Aeronautics and Space Administration. The fact that government generated data bases are heavily used is a function not only of their usefulness but also of the fact that their production and use are subsidized by the government.

Many of the large scientific, technical, and discipline oriented data bases have been produced by professional and technical societies in the not-for-profit part of the private sector. Some of these are: the SPIN (Searchable Physics Information Notices) tapes of the American Institute of Physics; BA Previews (Biological Abstracts Previews) of BioSciences Information Service; CA Condensates of Chemical Abstracts Service; PATELL (Psychological Abstracts Tape Edition-Leased or Licensing) of the American Psychological Association;

COMPENDEX (Computerized Engineering Index) of Engineering Index, Inc.; and METADEX (Metals Abstracts Index) of the American Society for Metals. These data bases are produced within the private sector, however, many of them have had research and development funds from the government which helped them to get started or to conduct research associated with systems or products.

The number of for-profit organizations producing data bases is small but some of the data bases are very important. For example: the Institute for Scientific Information publishes the Science Citation Index (SCI) tapes and the Social Science Citation Index (SSCI) tapes; Excerpta Medica is produced by the Excerpta Medica Foundation; the F & S Index of Corporations and Industries is produced by Predicasts, Inc.; and the New York Times Information Bank is produced by the New York Times.

These are but a few of the machine-readable data bases in use today. They are generated by government, for-profit and not-for-profit organizations and their orientation may vary. They may be discipline oriented, mission oriented, problem oriented, inter-disciplinary or multi-disciplinary. There are many of them and the level of use is rising rapidly. The data bases may be processed by centers that provide services directly to users, or services may be provided indirectly through brokers or service centers. The searches may be conducted on-line or in the batch mode.

Data base searching has a direct impact on libraries in several ways: it may affect the acquisition policy of the library--either increasing or decreasing acquisitions by either pointing out the non-use of some journals or the need for other journals; it may affect the interlibrary loan traffic of the library as either a borrowing organization or as a lending organization--depending on the

correspondence between the library's serials and monograph collections and the retrieved citations from data base searches; the library may expand or deepen its services by offering data base search services from data bases it processes; it may offer data base services to its clientele by functioning as an intermediary preparing search questions and processing them via an on-line service or through another center; or the library may function as a referral center directing its customers to the appropriate data bases and service centers.

Data bases relate to the National Program not only with respect to the service aspect, which bears directly on individual libraries, but also in areas where NCLIS has specifically expressed concern. There is a need for training programs to prepare librarians and information scientists to work with data bases. There is a need for continued federally supported information science and communication science research which would affect data base use. There is a need to expand data base content to serve new constituencies. There is a need to recognize and cooperate with the private sector in the generation and use of data bases. There is a need for development of resource locators and document delivery systems for closing the information retrieval (data base retrieval) loop. There is a need for working towards a reduced rate for telecommunications, (for information transfer) in order to promote and expand data base use and provide service to a wider range of users. And, above all, there is a need for data base sharing via networks--including all sectors of the information community.

Data Base Origins and Generation

Why were so many machine-readable data bases produced in the late 1960's and early 1970's? A data base exists once a file has been converted to machine-readable form. A few data bases were generated specifically for the purpose of information retrieval, but because the cost of data input is high and could seldom be justified for purposes of retrieval alone many more data bases were created as by-products of other activities. Some were created because machine-readable data was needed as a component of a computerized process control or production system for publishing primary journals, indexes or abstracting journals. Others were created as a result of the fact that computerized typesetting was used to produce a hard-copy publication. Computers have proved to be economic and effective tools for producing primary and secondary publications. Consequently, every time a publisher uses photocomposition a potentially machine-searchable file exists. The machine-readable file, once created, can be automatically reorganized, merged with other machine-readable files, reformatted, and repackaged to meet the demands of various markets. It is obvious that machine-readable files are considerably more flexible and can serve many more functions than can hard-copy records.

The by-product aspect of machine-readable data bases is no longer the raison d'être for many of today's major data bases. Many publishers or organizations in the business of information handling have adopted a "data base approach" to management of their processing systems and distribution of their information files. In such organizations the data base management system impacts all the information functions of the organization--abstracting, editing, indexing, authority files, production schedules, sequencing of tasks etc.--through to the composition and distribution of products whether printed,

microform, or machine-readable. "The data base approach, [which has been adopted by many publishers] asserts that there exists, for each enterprise, an accumulation of information that is pivotal to its operation. This concept implies that the description and treatment of such a collection should not be oriented toward specific processes but should be determined by the value and character of the information itself. An integrated data base usually means an organized collection of computer-readable information in which the information about each entity is recorded once in standardized form, and all access to that information is achieved through indexes and cross-references to the basic record and the authority files that support it." (Ref. No. 1) An integrated data base management system then, requires definition, design and standardization of the data elements that comprise the files.

Technical Aspects--Data Base Formats and File Structures

Before discussing data base searching and search services, a few distinctions are in order. It is helpful to understand the makeup of a data base in terms of the records and data elements that comprise it; the difference between the distribution format of a data base as it is provided by the producer and the format of a data base as the processor has structured it for searching; the meaning of batch, on-line and interactive; and, of course, the difference between retrospective and current awareness searching. Data elements are the basic building blocks of data bases. In the case of bibliographic data bases some of the generic names of the elements may be author, title, journal name, volume number, issue number, date of publication, index term, keyword, abstract, publisher name, etc. The data elements are the smallest units or elements that comprise the records (in this case bibliographic records) which, in turn,

comprise the file. A searcher is permitted to access individual records within the file or individual elements within the records. Thus, one can require specific index terms in a search question; the computer searches the index portion of the records in the data base to locate term matches and then produces the records that match the question. On the other hand, if the searcher knows the citation or reference number of certain desired records he may specify these and the records will be printed out or displayed.

It is possible to search specified data elements within an entire record either because the elements are identified or tagged with unique codes, or because the position of an element within a record may specify the type of element it is. Often a directory is associated with each record and it specifies the elements that are present, their location in the record, and the length (number of alphanumeric characters) of the data content of the data element. The standard arrangement of data element tags, data content, and directory information for the records is referred to as the format of the record, and the arrangement of the records on a tape or other media is referred to as the file structure or file format.

Unfortunately, file formats and record formats are not standardized nor are the definitions, contents, and representations of the data elements. Librarians tend to define data elements functionally as seen in the MARC II format of the Library of Congress, while information scientists generally define elements on the basis of content. There are almost as many data base formats as there are data bases, which leads to some confusion and, of course, added expense in processing tapes because it requires the processer of multiple tapes to either develop multiple search programs, or to reformat all incoming

tapes to a standard format. One important standard has been developed by the American National Standards Institute Inc. (ANSI) for interchange (transmittal) of bibliographic records--the "American National Standard for Bibliographic Information Interchange on Magnetic Tape." The MARC implementation of this standard has been proposed as a Federal Information Processing Standard (FIPS) and, barring problems, it will go into effect as a federal standard in August 1975. This standard deals only with the format for records on tape or the generalized structure but not the contents of the records. It does not define data elements or tags, specify required data elements, or specify data representation beyond that of the required character set.

Most data bases are distributed by their producers to processors in the form of sequentially arranged records on magnetic tape. The processors may either search the file in the distribution format or they may reformat the data base and store it on tape, disc or other media for searching. There are conceptually two basic structures for searchable files--sequential and inverted. (In fact, there are other forms or physical representations of the basic structures, e.g., direct access or index-sequential. There are also many ways of using several structures for different parts of the same data base.) In a sequential file records are arranged in sequence with all of the elements for a given record retained in one place and identified by the record ID. In an inverted file the searchable data elements are sorted with all postings (record ID's) that pertain to a given entry (e.g., an index term or title term) associated with that entry. Thus, for example, all ID's for records containing the term "CANCER" would be posted to that term. When an inverted file is used for searching, certain designated elements may be inverted while other elements remain with the full

bibliographic record file which is used for producing the output. Elements such as author names, patent numbers, and key words are useful search terms and so are inverted. But other elements are not inverted because they would be of little or no value as search terms. Page numbers, for example, are not inverted as it is unlikely a user would ask for all articles that begin on page 422.

An on-line system is one in which the user--through a terminal--is in direct communication with the central processing unit of the computer. An interactive system is one in which there is literally an interactive two-way communication between the user and the machine, and the time for response by the machine is, or should be, immediate. On-line searches of bibliographic data bases are usually run against inverted dictionary-type files. A batch processing system on the other hand, is one in which multiple jobs or search questions are "batched" together and run at one time. The search questions may or may not be entered via a terminal but they are saved until the time of the batch run. Searches against a serially or sequentially arranged file are usually run in the batch mode because the basic cost of spinning the tape once can be spread over several search questions rather than requiring one question to bear the total cost. There is, of course, some incremental cost for processing the additional questions.

Retrospective and current awareness searches differ with respect to the currentness of the files against which they are processed, and with respect to the number of times the question is run against the files. A retrospective question is one which is run against older, historical or past files, whereas a current awareness search is run against only the current or most recent file.

A retrospective question is usually run once against a collection of many data base issues or volumes, while a current awareness profile is run many times--each time against a different issue of the data base. Computerized current awareness systems are usually called SDI (selective dissemination of information) systems. Information is searched for and selected from the file in accordance with the users search profile. The output or search results are disseminated to the user(s). In the case of SDI, once a profile of the users interests is developed and refined, it is run on a regular basis against new issues of the data base(s) requested by the user. SDI searches are usually run in the batch mode against sequential files. After the SDI run is completed, the tape is added to the retrospective file for the appropriate data base. Several of the on-lineservices now offer SDI in addition to retro-searching. Since they have to process the incoming new data base issues as they arrive in order to add them to the retrospective files,they can conduct the SDI searches at the time of that initial processing. In these cases search output can be disseminated to the user through the mail or stored for later retrieval through his terminal. Retrospective questions may be run in either the batch or on-line modes depending on the system where the question is processed. In most cases the file that is searched is in inverted form for fast searching.

SDI and retrospective searches differ in purpose. The purpose of a retro-search may be to provide the user with: a) a few relevant references to become acquainted with a topic; b) a thorough coverage of the literature on a particular subject; or c) one or more references that contain the answer to a specific question. These searches are conducted or demand and in "past" or retrospective files. The completeness of the search question processed against the file varies considerably with the users purpose. In contrast,SDI searches are

conducted in order to keep the user up to date with the published literature in his field. The user profile is usually designed to be as complete as possible and to achieve high recall. The same profile is used over and over against new issues of the data base. The profile is, of course, modified over the course of a year if changes in user interests or data base output indicate the need. Since SDI and retrospective searches of data bases differ in purpose, comparisons of the two with respect to performance and cost make little sense. They are different services.

## Data Base Characteristics--Criteria for Use

### Subject Content

There are many different data bases and their differences can be described in terms of their characterisitics. It is on the basis of various combinations of characteristics that a user or center decides to search or offer services from a given data base. Certainly the first and most important difference is that associated with the subject matter covered by the data base. A data base with appropriate coverage is needed to effect a proper match between the user question and the data base to be searched. As described earlier, data bases are discipline oriented, mission oriented, problem oriented, or even multi-diciplinary or inter-disciplinary in character. Examples of disciplinary data bases are CA (Chemical Abstracts) Condensates, Polymer Science and Technology (POST), PATELL (Psychological Abstracts Tape Edition Leased or Licensed) and MEDLARS (Medical Literature Analysis and Retrieval System). Examples of mission oriented data bases are the Nuclear Science Abstracts data base produced by the Atomic Energy Commission and the STAR (Scientific and Technical Aerospace Reports) data base of the National Aeronautics and Space Administration. Problem oriented data

bases are HEEP (Abstracts on Health Effects of Environmental Pollutants) and PIP (Pollution Information Project), a data base prepared by the National Science Library of Canada using input that is selected from several commercially available data bases. An inter-disciplinary data base would be CBAC (Chemical and Biological Activities) and two multi-disciplinary data bases are the Institute for Scientific Information's Science Citation Index (SCI) covering virtually all areas of science and technology, and the MARC (Machine-Readable Cataloging) tapes of the Library of Congress covering most of the monographic literature processed by the Library of Congress regardless of subject content.

Other characteristics of data bases that affect the quality, timeliness and thoroughness of search results and the cost of processing are: (a) the type of source material included (journal articles, monographs, reports, theses, government literature, critical reviews, book reviews, newspaper articles, patents, etc.); (b) completeness of coverage (cover-to-cover, selected articles, selected issues of a journal, "all" versus "selected items" of any type); (c) lapse time between the appearance of an item in the primary source, the secondary source, and the machine-readable data base (note that the machine-readable product may precede the printed secondary source); (d) indexing and coding practices employed (free language keywords, controlled and uncontrolled index terms, author titles versus augmented or edited titles, codes to indicate subject matter, types or classes of any sort, etc.); (e) availability of abstracts, extracts or text on the data base for search and/or display; (f) data elements included for search (access points) and/or display (author, title, journal references, index terms, codes, cited references, etc.); (g) size and growth rate (How many records or references are contained in the file from the first year of the data base through the last completed year? What is the size

of an average record in terms of number of characters? And, what is the percentage growth rate of the data base per year?); (h) frequency of issue or update (how often are new issues of the data base produced? --weekly, semi-monthly, monthly, bimonthly, monthly, quarterly, annually, etc.).

Another consideration is the data base's correspondence with hard-copy publications. Some data bases have a 1:1 correspondence with printed products, i.e., every record contained in the machine-readable form exists in the hard-copy counterpart. Some include all of the same references but without the abstracts. In some cases the data base is a subset version of the hard-copy form and in other cases the reverse is true. A few data bases exist only in machine-readable form hence, it is not possible to check search results against a hard-copy.

Data Base Search Services

What services are provided from data bases and who provides them? The services most often provided are SDI and retrospective searches in either the on-line or batch mode. An additional data base service offered by a few organizations is a private library service. A service that is related to data base searching but is, lamentably, seldom offered by the centers that process data bases is that of document delivery.

A private library service is one that permits the user (individual or company) to create his own machine-readable file either by designating that output from his SDI runs be stored on his own tape or disc file, or by specifying other records (e.g., company reports or items selected via his own library searches) he would like to have entered into his file. He may have his SDI output saved for several weeks or months until he wants to look at it

and then he can indicate which of the records should be retained for his subsequent use. The advantage of the system lies in the fact that every record in the users file represents his own judgement. The file is personally tailored and it is under his control.

Closing the information retrieval loop requires delivery to the user of relevant documents identified through a data base search. In many ways it seems that the search itself is the easier or at least less time-consuming part of the information retrieval task. All too often a searcher completes a successful search only to be frustrated by the inability to readily obtain needed documents. The process of document delivery includes two major components: the identification of the source location of the document, and acquisition of the document. Delays associated with either or both occur often. One may ask, what are or could be the roles and responsibilities of the A & I (abstracting and indexing) services (the major data base producers), information centers, and libraries in the document delivery process? Very few data base processing centers handle document acquisition. Generally it is left to the user to go to his library to obtain copies or inter-library loan use of a document. Closer ties between the organization that processes the data base and the library that locates and orders the document might simplify the problem. This has been done at Ohio State University's Mechanized Information Center where document requisitions are handled by the center. What the user really needs is not necessarily to be able to acquire the document immediately following completion of a data base search, he needs to be able to by-pass all the intervening activities, time lags, etc. involved in locating the source and ordering the document. During 1975 two data base producers, ISI and NTIS, simplified the document acquisition

process for documents cited in their own data bases. In the case of ISI, on-line users of their data base through Lockheed or Systems Development Corporation (SDC) are now able to use the accession number to order full text copies of relevant items through ISI's Original Article Tear Sheet service (OATS). A specific command is provided for the searcher to order copies of desired items directly on-line through the system. The article order is subsequently transmitted to ISI headquarters in Philadelphia where the orders are filled. Ordered items are mailed to the requestor within 24 hours. A similar capability is available for on-line ordering of NTIS documents.

Possibilities for solving the document delivery problem include: (1) the data base producer's maintaining copies of all documents cited in his data base, as done by ISI and NTIS; (2) the developing for on-line searching of one or several union lists of serials with holdings information. Lists could be prepared on a national, state, or regional basis; (3) a national serials resources center functioning as a central depository. Any one of these solutions would simplify the problem of knowing where to find the document. The subsequent ordering can be simplified because we do in fact have a nation-wide communication network. The actual reproduction and delivery of the document is a separate problem. ISI's solution via use of tear sheet copies is certainly a good one and takes care of the copyright problem. The more common solution--the use of copying equipment--appears to be the easiest way of producing copies, though the legality still remains ambiguous. Facsimile transmission is used in cases where fast delivery is mandatory but this technique is still very expensive. However, if the National Commission were able to obtain lower communication rates for information transmission the expense would be reduced. The most inexpensive means of transmitting copies is still the U. S. Mail Service.

"SDI is one of the most successful information services developed in the past decade. A number of factors have led to the development of SDI: increased availability of computers; the automatic generation of data bases through computer typesetting; expansion of the worlds literature; and the increasing cost of labor-intensive information services." (Ref. No. 2) SDI has also been the primary use to which data bases were put in the early years of data base development. The reason that computer searching of retrospective data bases did not come into its own until the middle 1970s' is largely because only a few data bases with a sufficient number of years worth of material to make retro-searching worthwhile existed. A weekly or monthly SDI service from a data base during its first year is useful, but a search of a 5-month or 10-month collection of a data base is not very useful for retrospective search purposes. The situation has changed in the past year or two. The use of on-line search services for retrospective searching has grown by leaps and bounds. The number of on-line retrosearches conducted in 1974 has been estimated to be 700,000 (excluding library function uses of cataloging data files, e.g., OCLC) and the figure is likely to be 1,000,000 in 1975. (Ref. No.3) Reasons for this fast take-hold of on-line services are: user familiarity with other types of on-line systems such as airline space location systems; the availability of a sufficient number of years worth of data base cumulations to make retrospective searching useful; the relatively low cost of on-line searches; and, user familiarity of data bases via prior use of SDI services.

Data base search services may be used directly by the end-user or indirectly through centers or brokers. The use of the term "center" refers to organizations that acquire and process data bases themselves and provide services to users who may be within their own organization or outside of it. The term "broker"

refers to a person or organization that searches data bases on-line at another location or purchases data base searches from another center, for its own customers. The broker does not process the data bases but does provide search services from them. Obviously, the use of a center or broker involves some additional cost. The added cost for the middleman must then be offset by the added value provided by the middleman. The added value may consist of:

- augmentation, analysis, screening or interpretation of output
- know-how in effectively using other search services
- knowing where to go to find the appropriate service or data base
- document location and delivery
- reduction of the purchasers need for additional personnel with specific skills where the need for such skills may be sporadic
- reduction of the purchasers need for equipment, e.g., terminals, readers, etc. where the frequency of use is not sufficient to support the equipment

Data base processing centers may be independent commercial organizations, they may exist in computer centers, libraries or information centers of various sorts. More often than not, the processing of data bases has been done outside of the library setting. The brokerage situation though is different, because the brokerage organization needs little investment for equipment and has no need for programmers and computer specialists. Reference librarians or information specialists hired by libraries or information scientists operating independently can establish a search service for users with little capitol investment. They can effectively function as intermediaries between users and the systems.

Intermediaries

The computer is a tool that can greatly assist and speed up human activities but it is not a substitute for intellectual activities. In information retrieval the intellectual aspects of searching remain the prerogative of the searcher while the repetitive, routine and non-intellectual tasks are handled quickly and effectively by the computer.

Centers that operate computer-based information services as well as organizations that make use of on-line information services, in general, provide an intermediary between a user and the computer-based system. The intermediary may be an information specialist, information scientist or reference librarian. He or she handles the intellectual tasks of: selecting the appropriate system and data base(s) for the user's question; negotiating the search question with the user; developing the query or profile with an effective search strategy; conducting the search; and possibly evaluating the output. Additionally, the intermediary can not only maintain familiarity and expertise with systems and their various features, data bases and vocabularies but he/she can keep up with the changes that are made in systems, data bases and vocabularies.

Beyond the benefits that accrue to the use of intermediaries for handling data base searches, there are advantages associated with centralizing data base search activities within an organization. Advantages include: (1) use of knowledgable intermediaries for effective searching; (2) minimizing the number of personnel needed for data base searching; (3) distribution of the search personnel costs over a wider base, the development and use of one system for record keeping associated with searches; (4) developing in one location the personnel with the ability to negotiate contracts for data base activities; and

(5) minimizing the number of contracts negotiated by the organization. An intermediary may function as a searcher or in a referral capacity. In either case he/she must be aware of a multiplicity of sources and services and this is not a simple task.

## Data Bases and the National Program

There are many ways in which the National Program relates to data bases and these are in areas where NCLIS has expressed concern such as: education and training; research; constituencies served; generation and use of data bases within the private sector; resource locators; telecommunications; and network resource sharing.

### Education and Training

In line with NCLIS objective to "Develop and continually educate the human resources required to implement a National Program..." (Ref. No.4, p. 60) there is a real need for both basic and continuing education to train personnel to become skilled in the processing and searching of machine-readable data bases. Courses dealing with data base preparation, processing and use as well as courses in center operation and management are needed in the library and information science schools. There is a crying need for information specialists trained in modern techniques for information retrieval. "The spread of...modern methods of retrieving information is not a disease-like process but rather a force that makes it possible for the scientist to become society's eyes and ears--its overt intelligence service." .(Ref. No. 5) Scientists and the general public who have information needs must be willing and able to define what they need to know.

Likewise, information specialists, brokers, and information service librarians must know what resources (whether hardcopy, machine-readable or micrographic in form) and services exist and are appropriate for the users' needs; and then either refer the user to the proper source, or be able themselves to translate those needs into search questions and conduct the search.

Unfortunately, today "...most of our educational institutions are not turning out professionals who are technically equipped to deal with non-print materials or with computer and communication technologies." (Ref. No. 4, p. 51) It would certainly be beneficial to the information community as a whole, for NCLIS to promote, foster and fund specialized courses and programs in the use of modern techniques for information retrieval.

Research

The commission recognizes the role that the Office of Science Information Service of the National Science Foundation (OSIS/NSF) plays as "...the principal component of government responsible for information science research." (Ref. No.4, p. 85) OSIS/NSF has certainly played a significant role in the data base research and development (R & D) field. It has been responsible for R & D associated with design, management, and use of highly sophisticated data bases such as the Chemical Abstracts Services' Registry System. It was responsible for development of methodologies for analysis, use and evaluation of a wide variety of data bases and data base services through sponsoring the design and development of the university based centers at the University of Georgia, IIT Research Institute, University of Pittsburgh, UCLA, Ohio State University, Lehigh Univeristy, and Northeast Academic Science Information Centers.

Results of several OSIS/NSF current research grants will greatly impact on the data base field with respect to facilitating use of data bases and resource sharing of data bases through networks. MIT is investigating the feasibility of accessing multiple on-line systems and data bases through a common query language and common vocabulary. (Ref. No. 6) EDUCOM is conducting a sophisticated gaming and modeling study in order to investigate networking with respect to the economic, administrative and managerial aspects of various network configurations. (Ref. No. 7)

The University of Illinois is investigating the feasibility and utility of data base mapping via the interconnections (commonality of data elements and subject content) and potential routes that exist from one data base to another. The existence and location of data base resources together with an indication of the sequence in which they should be accessed will be shown. (Ref. No. 8) This NSF sponsored program should help pave the way to meet the need expressed by NCLIS: "Much of the success of a nationwide program will depend on knowing what information is available where, and how to gain access to it." (Ref. No. 4, p. 98)

The System Development Corporation is studying the impact of on-line search services (Ref. No. 9) and Lockheed Missiles and Space Company, Inc. is conducting an experiment in providing on-line reference retrieval services to the general public through public libraries. (Ref. No. 10)

NSF sponsored research at CAS has certainly resulted in some of the most significant developments in the field of chemical information, e.g., the Registry System for identification of unique chemical compounds and the development of schemes for substructure searching, automatic naming of compounds from structures, and automatic development of structures from names. OSIS/NSF has also sponsored work leading toward the reduction of duplicate efforts among data

base producers (Ref. Nos. 11 & 12), the use of common or standardized data element on machine-readable files, and, in general, has effectively encouraged cooperation among data base generators and among centers that process data bases. The commission would like to see a strengthening of the OSIS/NSF research and development programs and this should be encouraged. The existence of the technological feasibility of a national network and the advancement achieved in the state-of-the-art of data bases would not be where they are today if it had not been for research and development sponsored by the OSIS and the Division of Computational Research of NSF.

Constituencies Served

The use of machine-readable data bases relates to the Commission's objective to "...provide adequate special services to special constituencies, including the unserved." (Ref. No. 4, p. 55) There are efforts underway to develop new specialized data bases dealing with the problems of everyday life such as consumer affairs, legal aid, day care centers, recreational, health, and social services, etc. Such data bases may be developed for, and would be used by, neighborhood or community information centers. In this way the depth of service to current users would be expanded and services would be expanded to include new clienteles and new constituencies.

Resource Locators

One of the chief obstacles to wide use of available data bases and to the sharing of resources is the lack of public knowledge about the existence and location of available resources whether they exist within the federal government, state governments or the private sector.

There are aids such as: author addresses provided in many primary journals and in ISI's Current Contents for facilitating the readers ability to request reprints from authors; simple article ordering services such as ISI's original article tear sheet (OATS) service and the document ordering post cards bound in a number of journals requiring the requestor to merely circle the ID number of desired items; and, more recently, the provision of ordering from specific data base producers via on-line search services. These services are good but they cover only a limited number of resources. Other resource locators are needed.

The National Program can be instrumental in promoting and funding the development of tools for locating resources--data bases, search services, and back up documents. The location of data bases and search services will be greatly assisted by the use of tools currently being developed within the private sector by the not-for-profit organizations. Several surveys of sectors of the data base community are underway via the American Society for Information Science (ASIS), the Association of Scientific Information Dissemination Centers (ASIDIC), the National Federation for Abstracting and Indexing Services and at the University of Illinois (Ref. No. 8). The results of these efforts could be made more widely available for network use via the National Program. Similarly, but on a much larger scale, there is a need for a United States union list of serials holdings on-line (as has been done in Canada) for location of hard-copy documents to complement data base searching and complete the information retrieval loop. The development of this data base related tool and the network for accessing it would certainly be within the scope of the National Program.

Telecommunications

The continued and expanded use of machine-readable data bases is largely dependent on telecommunications and, as on-line use of data bases grows, the dependence will increase. The existence of common carrier communications networks, such as TYMNET , has been instrumental in the development of on-line data base services, however, the cost associated with communications has also been a barrier to some information services. At present communication charges represent approximately 10-20% of the out-of-pocket on-line search charges. The number, of course, varies with the data base accessed and the users location with respect to the computer site where the data base is searched. If the commission were to effect a lower tariff rate for information transmission it would certainly promote increased remote use of data bases, sharing of resources, and the use of networks for resource location. It would also promote the use of facsimile transmission for communication of information such as document requests or document delivery, as a result of the data base searches. Communication costs have long been a barrier to the use of facsimile transmission.

Although there is today no national program or plan for networking and resource sharing for machine-readable data bases, there is in fact a nation wide network over which data bases are shared. Specifically I refer to the use of the TYMNET communications network for searching data bases via remote terminals by many simultaneous users who are located virtually everywhere throughout the country (in fact several non-U.S. countries use the same facilities).

Communication satellites are now operating in the U. S. and internationally. This important development enlarges the nation's capability for exchange of information in all forms and a nation wide information network as proposed by

NCLIS, will need "...to integrate teletype, audio, digital, and video signals into a single system." (Ref. No. 4, p. 82).

Resource Sharing and Networking

The number of data bases, size of data bases and associated costs of operation provide the economical necessity for data base resource sharing. Any discussion regarding the need for data base services today and in the future necessarily involves a discussion of the reasons for, and advantages of, data base resource sharing and networking. Although data base sharing can be effected in many ways, the principal way in which sharing takes place today is by remote accessing of data bases through communications networks. No data base processing center whether it exists in an academic, industrial, or governmental orgainzation or whether it functions through the computer center, information center, or library can afford to process and provide services from all of the available data bases. Data base generation is expensive and so the costs of production, which are passed on to organizations that process data bases (for on-line or batch searching), are substantial. The cost of establishing and maintaining processing/searching activities is also high as it involves considerable investment in: data base purchase/lease/licensing; data base royalty and access fees; materials and equipment; machine time; communications; and personnel expenses. Additionally, the cost of preparing, negotiating, and conducting searches is high.

The principal advantages of data base resource sharing and networking are: availability of resources to a much larger community; reduced cost of data base searches as a result of distributing fixed costs over a larger base; reduction of the number of skilled personnel needed for processing data bases; accumulation of a wider variety of experiences and "know how" in data base use; development

of impetus toward standardization of data base formats, element definitions, formats for search strategies, access procedures and protocols, etc; and availability of more resources at a single location; and availability of data bases that individual user organizations would be unlikely to process internally because of low demand within the organization.

While resource sharing is largely done via communications networks, and of on-line systems, other types of sharing exist. For example, centers that process data bases themselves and provide services to clients (internal and/or external) often require services for their own clients from data bases that are processed in other centers. In such cases, two centers may exchange services or sell services to each other. Centers that provide their own batch processing, SDI, or retrosearch services often function as middlemen in accessing on-line services for their clients. On the other hand, they may function in a referral capacity in directing clients to the appropriate source.

Data Base Generation and Use in the Private Sector

NCLIS recognizes the necessity and advantages of accomodating, in the National Program, the wide range of resources and services within the private sector. They are an important part of the total information supply system today and will continue to be in the future.

One of the major areas where the National Program relates to data bases is the private sector which includes: the publishing industry; abstracting and indexing services, many of whom produce machine-readable data bases; the information industry and audio visual industry; and the special libraries in business and industry, many of whom are users of data base products and services.

Additional members of the private sector that affect data base activities directly or indirectly are the manufacturers of computers, terminals, user communication equipment, and the operators of communications networks such as TYMNET.

While access to government generated data bases such as ERIC, MEDLINE, CAIN, and NTIS is important, it is equally important that researchers and information service seekers in general have access to the New York Times Information Bank, Chemical Abstracts Condensates, The American Institute of Physics' SPIN, Engineering Index's COMPENDEX, PREDICASTS, and many others. Coordinated access to all of these sources in both sectors is needed in order to provide the types and levels of service required by users. The economic viability of data bases in the private sector is obviously related to level of use and rates charged for selling, leasing, licensing, and accessing the data bases through second and third party users (processors and brokers). The private sector recognizes and appreciates the fact that either directly or indirectly they have benefited from the government's involvment in the data base field. The government has subsidized many R & D programs in the private sector associated with systems and products. It has also funded the initial planning and development of centers that process data bases and sell service to users. All of this has been instrumental in bringing the data base "industry" to its current position. The private sector is mindful however, that in some instances the government has taken actions that may have a negative impact on the private sector by way of intervention, and competition. Just as no one wants to kill the goose that layed the golden egg, it also seems unreasonable for the goose to kill its offspring.

The data base service area is a new one and unfortunately it is not always the case that the implications of actions taken unilaterally on the part of either sector are fully understood at the time they are taken. NCLIS should be mindful of the interfaces, interdependencies and separate responsibilities of the two sectors in developing the National Program.

Data Base Problems and Future Trends

The major data base problems are not technical ones. They are legal, political and psychological and are associated with a lack of national leadership, cooperative resource sharing, network arrangements, competition, marketing, copyright, standards, and continued economic viability. Hopefully, NCLIS will be instrumental in solving some of these problems.

There are strong indications that in the future: we will see more data bases, covering more subject areas, with more special purpose subset and merged data bases being developed; the volume of data base use will increase and the user clientele served will represent more diverse constituencies; more data bases will be made available on-line through networks and a larger share of the total data base use will be on-line; there will be more involvement of librarians in data base services and services will be made available through public librarians as well as in the academic and industrial organizations; the techniques of computational linguistics, automatic content analysis, and pattern recognition will be employed on a larger scale; there will be more emphasis on the man-machine interaction; and, systems will become easier to use through natural language communication.

## REFERENCES


1. HUFFENBERGER, M.A.; WIGINGTON, R.L. "Chemical Abstracts Service Approach to Management of Large Data Bases." Journal of Chemical Information and Computer Sciences, 15:1 (February 1975) 43.

2. WILLIAMS, M.E. "Use of Machine-Readable Data Bases." Chapter 7 in: Annual Review of Information Science and Technology. Volume 9. Edited by Carlos A. Cuadra and Ann W. Luke. American Society for Information Science, Washington, D.C., 1974, p. 239.

3. WILLIAMS, M.E. NEWSIDIC--Information Bulletin of EUSIDIC (European Association of Scientific Information Dissemination Centres). Issue number 4 (October 1974) 10-11.

4. NATIONAL COMMISSION ON LIBRARIES AND INFORMATION SCIENCE. "A National Proposal for Library and Information Services." 2nd Draft. September 15, 1974.

5. GARFIELD, E. Current Abstracts of Chemistry and Index Chemicus. (Number 560).

6. MASSACHUSETTS INSTITUTE OF TECHNOLOGY (MIT). "Research in the Coupling of Interactive Information Systems." Principal Investigator, Francis J. Rientjes. (NSF Grant No. SIS 74-18165).

7. EDUCOM (Interuniversity Communications Council, Inc.) "Simulation and Gaming Project for Inter Institutional Computer Networking." Principal Investigator, James C. Emery.

8. UNIVERSITY OF ILLINOIS, INFORMATION RETRIEVAL RESEARCH LABORATORY. "Data Base Mapping Model and Search Scheme." Principal Investigator, Martha E. Williams. (NSF Grant No. SIS 74-18558).

9. SYSTEMS DEVELOPMENT CORPORATION (SDC). "An Analysis of Man-Machine System Communication and Existing On-line Retrieval Systems." Principal Investigator, Carlos A. Cuadra (NSF Grant No. SIS 74-03465).

10. LOCKHEED MISSILES AND SPACE COMPANY. "Investigation of the Public Library System as a Linking Agent to Major Scientific, Educational, Social, and Environmental Data Bases." Principal Investigator, Oscar Firschein. (NSF Grant No. SIS 74-13972).

11. AMERICAN INSTITUTE OF PHYSICS AND ENGINEERING INDEX, INC. "Interchange of Data Bases." Principal Investigators, Rita G. Lerner and Robert H. Marks. (NSF Grant No. G.N. 42062).

12. BIOSCIENCES INFORMATION SERVICE AND CHEMICAL ABSTRACTS SERVICES. "A Project to Redesign and Re-engineer the BIOSIS System." Principal Investigator, Phyllis V. Parkens. (NSF Grant No. C-810).

## RELATED WORKS


ASLIB. Annual Conference, 47th, London, England, 23-26 September 1973. Proceedings.

BECKER, JOSEPH. The First Book of Information Science. U.S. Atomic Energy Commission--Office of Information Services, Oak Ridge, Tennessee, 1973, 91 pp.

FINER, RUTH, comp. A Guide to Selected Computer-Based Information Services. Aslib, London, England, 1972, 113 pp.

HENDERSON, MADELINE. Evaluation of Information Systems: A Selected Bibliography with Informative Abstracts. National Bureau of Standards, 1967. (ED 016 497)

HOLM, BART E., et al. "The Status of Chemical Information." Journal of Chemical Documentation, 13:4 (November 1973) pp. 171-183.

HOWERTON, PAUL W., ed. Management of Information Handling Systems. Hayden Book Company, Inc., Rochelle Park, 1974, 232 pp.

THE INTERNATIONAL DIRECTORY OF COMPUTER AND INFORMATION SYSTEM SERVICES 1974, THIRD EDITION. Gale Research Company, Detroit, Michigan, 1974, 636 pp.

KEENAN, STELLA, ed. Key Papers on the Use of Computer-Based Bibliographic Services. American Society for Information Science, Washington, D.C.; National Federation of Abstracting and Indexing Services, Philadelphia, Pennsylvania, 1973, 179 pp.

KEENAN, STELLA; ELLIOTT, MARNA. "World Inventory of Abstracting and Indexing Services." Special Libraries, 64:3 (March 1973) 145-150.

KRUZAS, ANTHONY T., comp. and ed. Encyclopedia of Information Systems and Services. International Edition. Edwards Brothers, Inc., Ann Arbor, Michigan, 1974, 1283 pp.

LANCASTER, F.W.; FAYEN, E.G. Information Retrieval On-Line. John Wiley & Sons, Inc., New York, New York, 1973, 597 pp.

LEE, CALVIN MARK, comp. An Inventory of Some English Language Secondary Information Services in Science and Technology. Organisation for Economic Co-operation and Development, Directorate for Scientific Affairs, Scientific and Technical Information Policy Group, Paris, France, June 1969, 21 pp. (Report No. DAS/STINFO/69.3).

MARRON, BEATRICE; FONG, ELIZABETH; FIFE, DENNIS W.; RANKIN, KIRK. A Study of Six University-Based Information Systems. National Bureau of Standards, Institute for Computer Sciences and Technology, Washington, D.C., June 1973, 96 pp. (NBS Technical Note 781).

MATHIES, M. LORRAINE; WATSON, PETER G. Computer-Based Reference Service. American Library Association, Chicago, Illinois, 1973, 214 pp.

MAUERHOFF, GEORG R. "Selective Dissemination of Information." In: Advances in Librarianship. Volume 4. Edited by Melvin Voigt. Academic Press, New York, New York, 1974, 25-62

SCHIPMA, PETER B. "Generation and Uses of Machine-Readable Data Bases." In: Annual Review of Information Science and Technology, Volume 10. Edited by Carlos A. Cuadra and Ann W. Luke. American Society of Information Science, Washington, D.C., 1975. (Not yet published).

SCHNEIDER, JOHN H.; GECHMAN, MARVIN; FURTH, STEPHEN E., eds. Survey of Commercially Available Computer-Readable Bibliographic Data Bases. American Society for Information Science, Special Interest Group for Selective Dissemination of Information, Washington, D.C., January 1973, 184 pp. (ED 072 811. Available from American Society for Information Science, Washington, D.C.).

VEAL, DOUGLAS C. "Computer Techniques for Retrieval of Information from the Chemical Literature." In: Fortschritte der Chemischen Forschung (Topics in Current Chemistry). Volume 39. Edited by F. Boschke. Springer-Verlag, New York, New York, 1973, 65-89.

WILLIAMS, MARTHA E. "Use of Machine-Readable Data Bases." Chapter 7 in: Annual Review of Information Science and Technology. Volume 9. Edited by Carlos A. Cuadra and Ann W. Luke. American Society for Information Science, Washington, D.C., 1974, 221-284.

WILLIAMS, MARTHA E.; BRANDHORST, W.T., co-contributing eds. A column on "Data Bases." In: Bulletin of the American Society for Information Science, published monthly, First Issue June/July 1974.

WILLIAMS, MARTHA E.; STEWART, ALAN K. ASIDIC Survey of Information Center Services. IIT Research Institute, Chicago, Illinois, June 1972, 117 pp.